

Case-based Social Statistics II

CSSS 321

Professor: Mark S. Handcock

Solutions to Homework 8

Due Friday, June 6, 2002

Problems to be handed in:

- 1) This question returns to the HeadStart type program considered in Homework 6. This program aims to improve the job-prospects of participants. In seeking to determine just how influential the HeadStart training is, the review committee of a program has collected data over the previous year on the weekly income on participants, and the number of months they were trained within the program. The data are available on the course website under "Data".
 - a) Using the "Simple Linear Regression" command in Datatools, refit the model for income in terms of the length of training. Click the button that saves the residuals (into the variable var3).

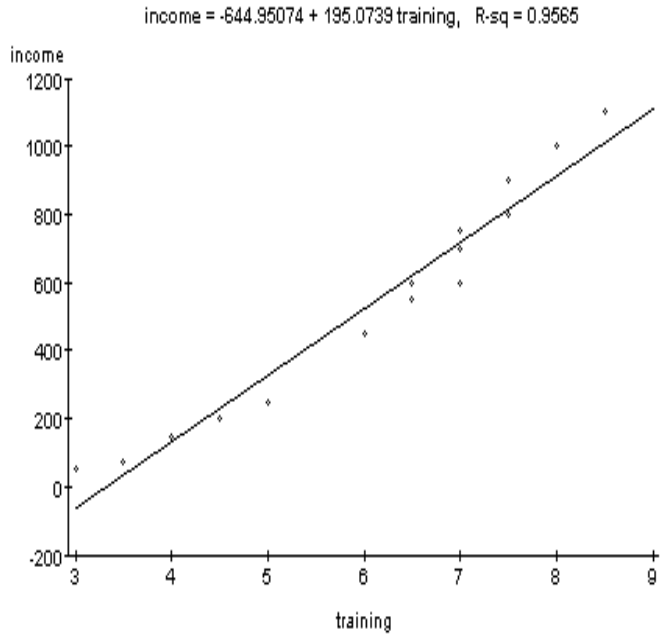
Solution(5): We now fit the simple linear regression model.

```
Commands:-----
Stat
  Simple Linear regression
    Dependent variable: income  Independent variable: training
Output:-----
Simple linear regression results:

Independent variable: training
Dependent variable: income
Sample size: 15
Correlation coefficient: 0.978
(See fitted line plot in Graphics Panel.)
Residuals stored in column var3

Estimate of sigma: 73.51859

Parameter   Estimate   Std. Err.   DF   Tstat     Pval
Intercept   -644.95074  72.89727   13   -8.847392  0
training    195.0739   11.538096  13   16.90694   0
```



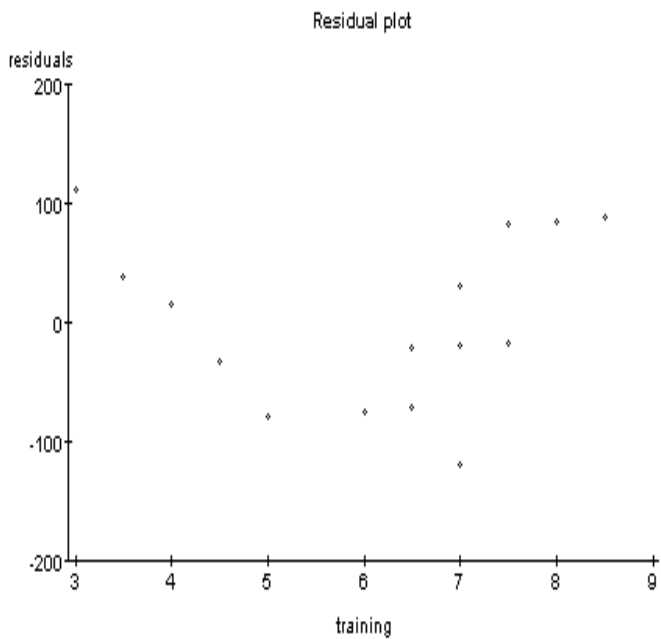
a) Using the “Scatter Plot” command, plot the residuals against the length of training.

Solution(1): In DataTools this is under the “Graphics” menu and then the “Scatterplot” command.

```

Commands:-----
Graphics
  Scatter plot
    training var3
Output:-----

```



- b) Interpret the Scatterplot. Do the four assumptions of linear regression appear to hold here?

Solution(1): With so few residuals it is difficult to make definitive statements. However, there does appear to be a pattern in the residuals with a clear “bowl” shape in evidence. This would suggest that the relationship is curvilinear rather than linear. No outliers exist, and normality and constant variance can not be directly assessed until the mean is corrected. They do not seem grossly incorrect though. A solution is to fit a multiple regression model with a quadratic term in training time. ✕

- 2) You run an apple farm in Yakima and obtain the following data containing information on crop yields, total rainfall and average temperature for this arid region of Washington over eight years.

variable	units	value							
-----	-----	-----	-----	-----	-----	-----	-----	-----	-----
time	year	1992	1993	1994	1995	1996	1997	1999	1999
yield	bushels/acre	60	50	70	70	80	50	60	40
rainfall	inches	8	10	11	10	9	9	12	11
temperature	fahrenheit	56	47	53	53	56	47	44	44

The data are available on the course website under “Data”.

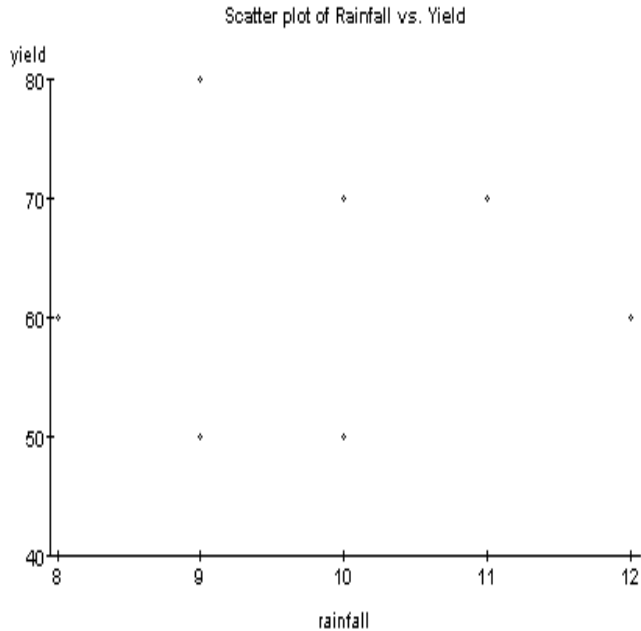
- a) Using the “Scatterplot” command in Datatools, plot the yield against rainfall, and yield against temperature. Does it appear that these variables are linearly related?

Solution(1): In DataTools this is under the “Graphics” menu and then the “Scatterplot” command.

```

Commands:-----
Graphics
  Scatter plot
    rainfall yield
Output:-----

```



- b) Using the “Multiple Linear Regression” command, fit a regression model for yield based on rainfall. Give an interpretation of the slope. Does it represent an intuitively reasonable model for the relationship between rainfall and yield. Why?

Solution(6): The first step is to read the data into DataTools.

```

Commands:-----
Stat
  Simple Linear regression
    Dependent variable: YIELD  Independent variable: RAINFALL
Output:-----
Analysis of variance table for multiple regression model:

Source   df    SS          MS          Fstat        Pval
Model    1    33.333332   33.333332   0.17142858   0.6932
Error    6   1166.6666   194.44444
Total    7   1200

Root MSE: 13.944334    R-squared: 0.0278

Parameter estimates:

Variable  DF    Estimate    Std. err.    Tstat        Pval
Intercept 1    76.666664   40.554604   1.8904552    0.1006
rainfall  1    -1.6666666  4.0253825  -0.41403934  0.6912

```

The regression line is then:

$$\text{yield} = 76.7 - 1.67 \times \text{rainfall}$$

That is, a one inch increase in rainfall leads to a 1.67 bushel/acre *decrease* in yield. As this is an arid region, increase in rainfall should increase the yield. This negative slope is not reasonable, and further investigation is necessary.

✕

- c) Using the “Multiple Linear Regression” command, regress yield on both rainfall and temperature. That is, fit the model

$$\text{YIELD} = \text{CONSTANT} + \text{RAINFALL} + \text{TEMP}$$

Note that the sign of the estimate for the slope of rainfall is opposite to that of (a). The vice-president of the company ask you to explain this phenomenon in simple terms. What do you say?

Solution(5): Returning to DataTools:

```

Commands:-----
Stat
  Multiple Linear regression
    Dependent variable: YIELD  Independent variable: RAINFALL TEMP
Output:-----
Analysis of variance table for multiple regression model:

Source   df    SS          MS          Fstat      Pval
Model    2    948.5714    474.2857    9.431818   0.0201
Error    5    251.42857   50.285713
Total    7    1200

Root MSE: 7.0912423    R-squared: 0.7905

Parameter estimates:

Variable   DF   Estimate   Std. err.   Tstat      Pval
Intercept  1   -144.7619   55.849895   -2.5919816  0.0411
rainfall   1    5.714286   2.6802375   2.1320071   0.077
temperature 1    2.952381   0.69203436  4.2662344   0.0053

```

The multiple regression line is then:

$$\text{yield} = -144.6 + 5.71 \times \text{rainfall} + 2.95 \times \text{temperature}$$

That is, a one inch increase in rainfall leads to a 5.71 bushel/acre *increase* in yield. A one degree increase in temperature leads to a 2.95 bushel/acre *increase* in yield. Now an increase in rainfall increases the yield! The reason is that the second factor, temperature, leads to lower rainfall and higher yields. Thus as the temperature increases the yield drops, while the rainfall decreases. Thus, overall, we see a pattern of lower rainfall and increased yields, even though both are driven by increased temperatures. Put in another way, the direct effect of rainfall is to increase yield. The indirect effect of rainfall on yield through temperature is negative, in such a way as to make the overall effect of rainfall a decreasing yield.

⊗

- 3) If the only possible values of a explanatory variable (independent variable) are 1 and 0, then that variable is called a “indicator.” The ‘1’ usually represents the presence of some attribute and ‘0’ represents the absence of the attribute. The regression coefficient in such a case is usually interpreted as the added response associated with the presence of the attribute. In this problem we will explore the use of a indicator variable.

The set of values listed below are the price-per-pound (PRICE) for twenty beef steers. Also given are the actual total weight (SIZE) and the USDA grade (PRIME). This are

coded so that PRIME=0 for “choice grade” steers and PRIME=1 for “prime grade.”
 The data are available on the course website under “Data”.

PRICE	SIZE	PRIME
36.75	736	0
35.50	506	0
28.00	394	0
31.00	419	0
32.75	504	0
35.25	599	0
31.50	442	0
33.50	505	0
32.50	512	0
31.25	400	0
33.50	339	1
36.00	517	1
33.00	396	1
30.50	382	1
36.00	479	1
35.25	515	1
37.50	606	1
30.50	336	1
33.00	425	1
34.75	498	1

- a) Find the regression equation for PRICE on both SIZE and PRIME. Note the regression coefficient of PRIME and give an interpretation.

Solution(4): The first step is to read the data into DataTools The variables are named PRICE, SIZE, and PRIME. Some simple information about the data can be obtained from the “Stat”/“Summary Stats” menu.

```

Commands:-----
Summary statistics
Descriptive statistics
variables: PRICE SIZE PRIME
Output:-----
Summary Statistics:
Variable  n    Mean   Variance  Std. Dev.  Median  Range  Min  Max
price     20   33.4   6.114474  2.4727461  33.25  9.5  28  37.5
size      20  475.5  9486.895  97.40069   488.5  400  336  736
prime     20    0.5   0.2631579  0.51298916  0.5    1    0    1
    
```

The required regression is

```

Commands:-----
Stat
Multiple Linear regression
Dependent variable: PRICE Independent variable: SIZE PRIME
Output:-----
Analysis of variance table for multiple regression model:

Source  df  SS      MS      Fstat    Pval
Model   2   89.517944  44.758972  28.544138  0.0000
Error   17  26.657053  1.568062
Total   19  116.175
    
```

Root MSE: 1.2522228 R-squared: 0.7705

Parameter estimates:

Variable	DF	Estimate	Std. err.	Tstat	Pval
Intercept	1	21.645367	1.5896462	13.616468	0
size	1	0.022233672	0.003068637	7.2454553	0
prime	1	2.3650444	0.58263874	4.059195	7.0E-4

The fitted model is

$$\text{PRICE} = 21.645 + 0.022 \times \text{SIZE} + 2.365 \times \text{PRIME}$$

This suggests that each 100 pounds of SIZE adds 2.2 cents to the cost-per-pound. The coefficient 2.365 of the indicator variable PRIME suggests that the better-grade steers are worth 2.365 cents per pound more.

We note also that the coefficients of SIZE and PRIME are both statistically significantly different from zero. The p-value for testing if the overall model is significantly better than the model without SIZE and PRIME is $0.000 < 0.05$. Thus the over all model is significantly better than the model without SIZE and PRIME. \boxtimes

- b) Calculate a 95% confidence interval for the price-per-pound difference between choice and prime steaks of identical weight

Solution(3): The price-per-pound difference between choice and prime steaks of identical weight is β_1 . A 95% confidence interval for β_1 is

$$\begin{aligned}\widehat{\beta}_1 \pm t_{\alpha/2}(n - k - 1) \times \widehat{SE}(\widehat{\beta}_1) \\ = 2.365 \pm t_{0.025}(17) \times 0.583 \\ = (\$1.13, \$3.60)\end{aligned}$$

where $t_{0.025}(17) = 2.110$. The values for $\widehat{\beta}_1, n = 20, k = 2$ variables and $\widehat{SE}(\widehat{\beta}_1)$ can be obtained from the DataTools output. \boxtimes

- c) Test the hypothesis, 99% confidence, that there is no price-per-pound difference between choice and prime steaks of identical weight

Solution(3): This is a test of the hypothesis that β_1 is zero. The null and alternative hypotheses are:

$$H_0 : \beta = 0.0$$

$$H_1 : \beta \neq 0.0$$

respectively. It is identical to the test that PRIME makes a difference in the model. Thus we can test it based on the p-value for PRIME. As the p-value is $0.0008 < 0.01$, we can reject the null hypothesis with 99% confidence. \boxtimes